by PAUL E. PETERSON & WILLIAM G. HOWELL

Voucher Research Controversy

New looks at the New York City evaluation

"Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City," "Comment," and "Rejoinder"

By John Barnard, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin; "Comment" by Alan Krueger and Pei Zhu

Journal of the American Statistical Association, June 2003.

Another Look at the New York City School Voucher Experiment

By Alan Krueger and Pei Zhu

Presented at the National Press Club, April 2003.

ILLUSTRATION BY TIMOTHY COOKE

In *The Education Gap: Vouchers and Urban Schools* (Brookings, 2002), we and our colleagues reported that attending a private school had no discernible impact, positive or negative, on the test scores of non-African-American students participating in school voucher programs in Washington, D.C., New York City, and Dayton, Ohio. But after one, two, and three years in New York City, and after two years in Washington and Dayton, significantly positive impacts for African-Americans were observed.

Our results came from randomized field trials, which are generally thought to be the gold standard for research on human subjects. In such studies, subjects are randomly assigned to treatment and control groups by means of a lottery. In the best of worlds, researchers are able to collect information on the subjects' characteristics before the lottery begins, enabling them to confirm that the lottery, in fact, worked as intended. If the treatment and control groups are similar at the beginning of the study, any differences between the two groups that emerge over time can be attributed to the programmatic intervention—in the case at hand, using a voucher to switch from a public to a private school. The results reported in this article are thus to be understood as the difference in test scores between those students who used vouchers to attend a private school and those of their public school peers who would have used a voucher had they been offered one.

Despite the strength of our evaluation's design, the findings have not been without controversy. Specifically, two secondary analyses of the New York City data have recently been published, with widely diverging results. One study, conducted by a group of distinguished statisticians, John Barnard, Constantine Frangakis, Jennifer Hill, and Donald Rubin (hereinafter referred to as Barnard), has confirmed our first-year results but has been virtually ignored in the public media. The other, by Princeton economists Alan Krueger and Pei Zhu, has contradicted our results and twice received favorable coverage in the *New York Times*, where Krueger is an occasional columnist.

From the standpoint of pure inno-

vation and analytical rigor, Barnard has produced the more impressive piece. As befitting an article published in the nation's leading statistics journal, it introduces new statistical techniques to deal with problems that often emerge in randomized field trials: 1) missing data (for instance, not all students who initially joined the study participated in the follow-up testing sessions), and 2) noncompliance (some students, for example, refused the vouchers that were offered to them).

It remains to be seen whether the statisticians' proposed innovation becomes more widely used. At its current stage of technical development, it permits the examination of effects only after one year. Also, in using the technique, Barnard opted to restrict their analysis to those families with only one child participating in the voucher program.

Despite differences in statistical approach and in the selection of students to be included in the analysis, Barnard's findings are largely consistent with those we reported. While we estimated that, after one year, African-American students scored 7 percentile points higher on the math portion of the Iowa Test of Basic Skills than their peers in public schools, Barnard reports impacts of 6 percentile points for African-American students from low-performing public schools. (Almost all the African-American students came from schools with average test scores below the district mean; the few that did not had almost identical average impacts, but the number of available observations was too small to recover precise estimates.)

By contrast, Krueger and Zhu concluded, "The provision of vouchers in New York City probably had no more than a trivial effect on the average test performance of participating black students." This conclusion rests primarily on three methodological decisions that distinguish their research from both our study and that of Barnard:

Our results came from randomized field trials, which are generally thought to be the gold standard for research on human subjects.

- We and Barnard let the mother's ethnicity define the student's ethnicity, while Krueger and Zhu defined a student as African-American if either parent was African-American.

- We and Barnard considered the results for only those students in grades 1–4, almost all of whom took achievement tests before the lottery. This provided us with what are known as "baseline test scores" that can be used to obtain more precise estimates of program effects. By contrast, Krueger and Zhu also included a large number of kindergartners for whom no baseline test scores were available.

- We and Barnard always adjusted the data to account for students' baseline test scores in estimating our results. Krueger and Zhu, in their preferred results, as presented in their "Comment" on Barnard, exclude these baseline test scores.

All three of these alterations to the research strategy must be made in order to obtain results that differ substantially from those that we and Barnard obtained. Using any one or two of these different strategies does not generate appreciably different results.

How to Define African-American

Let's consider Krueger and Zhu's decision to classify students as African-American if either parent was African-American. Krueger and Zhu regard this decision as a key reason why they obtained results different from ours.

To understand the issue, bear in mind that because many of the students were very young, their ethnic backgrounds were ascertained from information provided in questionnaires filled out by the adults who accompanied them to the testing sessions. These adults were asked to report the ethnicity of the student's mother and, separately, the student's father. They could assign parents to one of nine categories, five of which are: Black/African-American (non-Hispanic); White (non-Hispanic); Puerto Rican; Dominican; and Other Hispanic. Classifying a child's ethnicity is usually straightforward, because both parents are of the same background. In cases where parents were not of the same ethnicity, we classified the child by the mother's ethnicity, simply because most children lived with their mothers, 74 percent of whom were single parents. Sixty-seven percent of the students lived with only their mother, compared with just 2 percent who lived with only their father. Mothers accompanied 84 percent of children to testing sessions; in 94 percent of the cases, the accompanying adult claimed to be a caretaker of the child.

Given the fact that these children tended to live with their mothers (and, often, not with their fathers), the decision to link the child's ethnicity to the mother's appears perfectly sensible. Alternatively, one might classify students as African-American only if both parents are African-American or if the child's primary parental caretaker (usually the mother, but on a few occasions the father) is African-American.

Eschewing these alternatives, Krueger and Zhu used a unique clas-

sification scheme. They identify students of mixed heritage as African-American as long as either the mother or the father is African-American. If the mother was white but the father was African-American, the child was defined as "black, non-Hispanic." Even if a child had a Hispanic mother and an African-American father, Krueger and Zhu still classified the child as "black, non-Hispanic." Unless one departs from the standard practice of using mutually exclusive categories, students could not be classified as Hispanic or white if either parent was African-American. Krueger and Zhu defend this classification scheme on the grounds that it is "symmetrical." But symmetry is hardly the word for a scheme that classifies Hispanics, whites, and African-Americans according to different principles.

Nevertheless, not much turns on how one defines a child's ethnicity. Regardless of one's definition, impacts after three years that range between 7 and 8 percentile points are observed for African-Americans in New York City (see Figure 1). If one classifies a student's ethnicity by the mother's (the approach we prefer), the effects are 8 percentile points; if one uses either the mother or the father (the approach favored by Krueger and Zhu) the effects are 7 percentile points, a result that is not significantly different from the one originally reported. By itself, altering the definition of a child's ethnicity provides no basis whatsoever for concluding that effects disappear.

It is problematic to include students in a study if you don't know what their achievement level was at the beginning.

Students without Baseline Test Scores

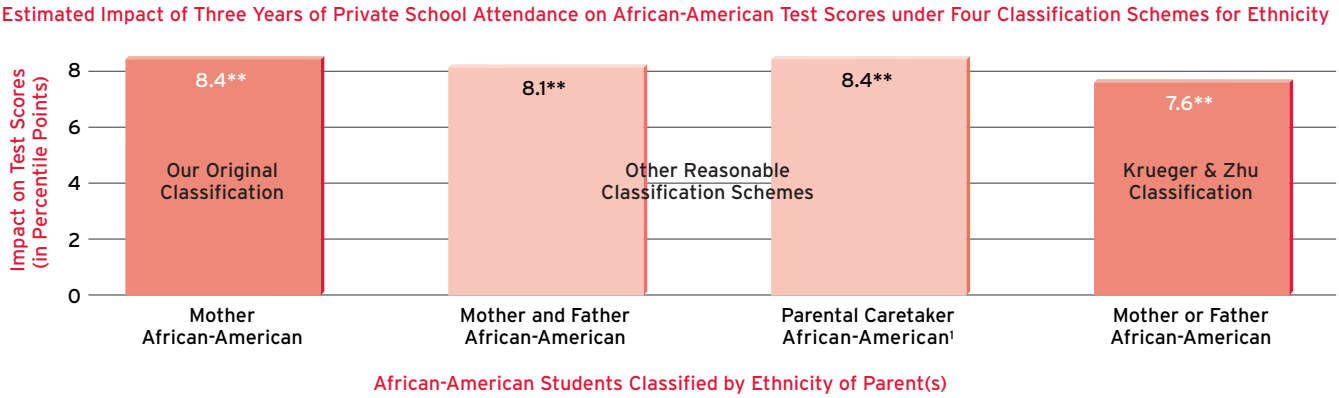
Figure 1 presents results for students with baseline test-score information—the first bar reporting impacts for the definition of African-American originally used, the latter three bars for alternative definitions. The figure's results are based on analyses that exclude from the study all kindergartners, none of whom were tested at baseline. Also excluded are the 10 percent of the students in grades 1–4 who were sick, who refused to take the test, or whose tests were lost in the administrative process.

Krueger and Zhu object to the exclusion of any students from the study, claiming that this constitutes the "most important" deficiency of our analysis, as well as that of Barnard. But even when all students are included in the analysis, African-American students who attended private schools scored significantly higher than their public school peers (see Figure 2).

Nonetheless, it is problematic to include students in a study if you don't know what their achievement level was at the beginning. How well students perform on a test at, say, age seven, is tightly connected to how well

Four Approaches, Same Result (Figure 1)

Whether one defines a student's ethnicity by his mother's (as we originally proposed) or by either the mother's or father's (as Krueger and Zhu have proposed), findings remain much the same. Two other ways of defining students' ethnicity also yield similar results.



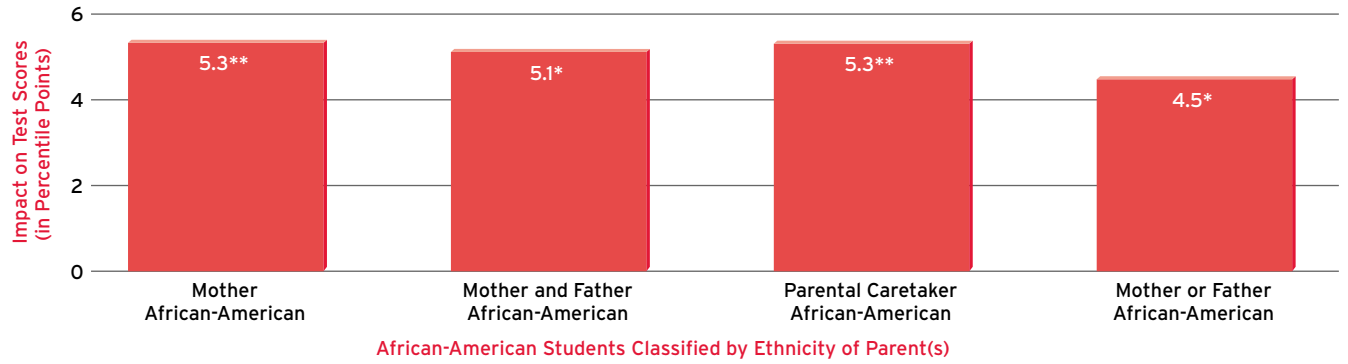
Note: Estimates are for students in grades 1-4 with baseline test scores.
[1] Mother assumed to be the primary caretaker of the child's education except in those cases where the child lives only with the father.
** Results significant at the .05 level.

SOURCE: Authors

Four More Approaches, Results Remain Significant (Figure 2)

When one includes kindergarten and other students without baseline scores (as Krueger and Zhu propose), private school effects attenuate somewhat but remain statistically and substantively significant.

Estimated Impact of Three Years of Private School Attendance on African-American Test Scores under Four Classification Schemes for Ethnicity



Note: Estimates are for students in grades K–4 with and without baseline scores, controlling for baseline test scores whenever possible.
* Results significant at the .10 level.
** Results significant at the .05 level.

SOURCE: Authors

they will do at age eight, nine, or ten. In fact, the correlations between baseline and follow-up test scores in New York consistently hover around 0.7. By comparison, the correlations between mother's level of education and follow-up scores were only about 0.1.

Restricting the study to those students for whom baseline test scores are available affords a check on whether the lottery worked as intended and whether any problems arose downstream. For these students, all looks fine on both accounts.

When including all students, even those lacking baseline test scores, one can only hope that the two groups are similar with respect to this critical characteristic. Nonetheless, Krueger and Zhu defend their inclusion on the grounds that "because assignment to treatment status was random . . . a simple comparison of means between treatments and controls without conditioning on baseline scores provides an unbiased estimate of the average treatment effect." This claim, says Barnard, "is simply false."

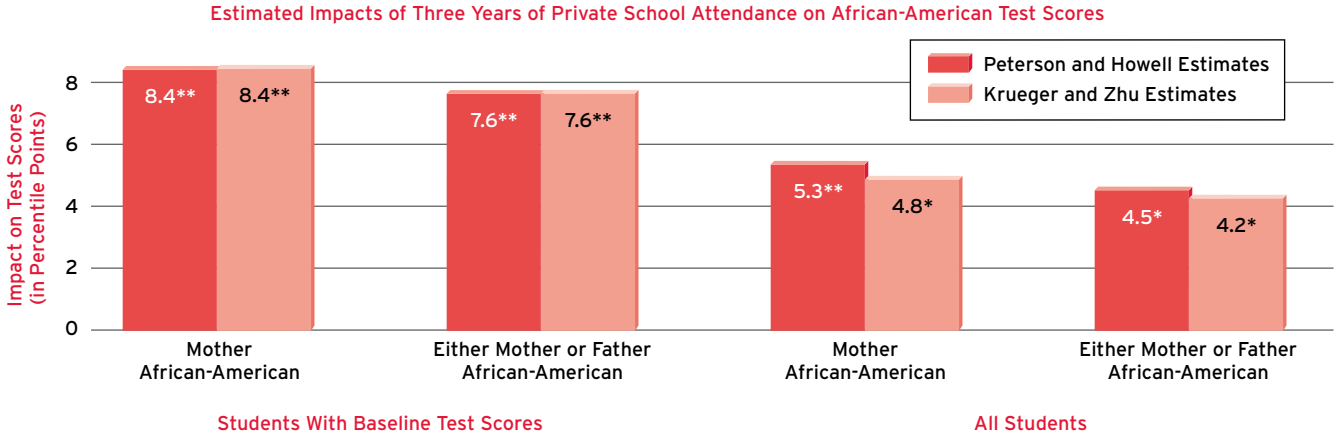
If not quite false, the claim is at least dubious, because there were many ways for the treatment and control groups to become unbalanced. For example, about a third of the students did not remain in the study into the third year—a fairly standard rate of attrition from this kind of research protocol, but one that raises concerns that the treatment and control groups might have lost students with different baseline test scores. For this reason, we limited our analysis to those students for whom baseline scores were available, and hence for whom we were able to verify that the treatment and control groups did not become unbalanced.

Neither changing the definition of African-American nor adding students for whom baseline test scores are missing appreciably changes the results we originally reported.

But perhaps something else is to be gained from including all students, regardless of whether baseline information was available. Krueger and Zhu suggest that by adding these cases one can generalize findings to another grade level (kindergartners). Unfortunately, this is a hazardous generalization, given the fact that the results for kindergartners were significantly different from those for the older students. African-American students in grades 1–4 scored significantly higher if they attended private school, a result observed in all three years of the study. The results for kindergartners, meanwhile, were considerably more erratic; the effect of attending a private school for three years was a negative 13.9 percentile points. In the absence of base-

Reconfirmed Results (Figure 3)

Krueger and Zhu's claim that they could not duplicate our results is misleading; the findings in their replication hardly differ from ours.



Note: Students with and without baseline scores, controlling for baseline test scores whenever possible.
* Results significant at the .10 level
** Results significant at the .05 level

SOURCE: Authors; Krueger and Zhu correspondence

line scores, we don't know whether the findings for kindergartners are genuine or simply the result of errors in the administrative process.

Krueger and Zhu also note that their inclusion of all students in the sample generates more precise estimates. But gains in precision obtained by increasing the number of students observed will be offset by losses associated with failing to control for baseline test scores. One can assess the extent to which these competing forces balance each other by comparing the estimates' standard errors: the smaller the errors, the more precise the estimate. As it turns out, the standard errors are larger, not smaller, when estimating statistical models that include all students but do not control for baseline test scores.

A compromise strategy, suggested by Krueger and Zhu, includes all students and adjusts for baseline test scores whenever possible. This analytic approach generates more precise estimates, the results from which are presented in Figure 2. But since these analyses also introduce risks of bias (principally by including the kindergartners for whom no baseline scores were available), the results in Figure 2 are inferior to the results provided in Figure 1. Nonetheless, they still reveal significantly positive effects of attending private schools on African-American test scores. In other words, even if one includes kindergartners in the study, as Krueger and Zhu recommend, the essentials of our original finding remain intact.

More than 25 years ago, James Coleman and his colleagues found that attending private schools was more beneficial for black students than for whites.

Krueger and Zhu have not accepted these findings, however. Instead, they have said that they cannot obtain equivalent results when they attempt to conduct an analysis identical to ours. But this claim is misleading. In fact, Krueger and Zhu's results, available by correspondence, hardly differ from ours. As the first set of columns in Figure 3 shows, among students with baseline test scores, we both find that the estimated year three private school impact is 8.4 percentile points for all African-Americans (as defined by the mother's ethnicity, our preferred definition). And, as shown in the second set of columns in Figure 3, we both find an impact of 7.6 percentile points for African-Americans when using Krueger and Zhu's preferred definition of African-American (students whose mother or father is African-American). Moreover, when students without baseline scores are added to the analysis, they obtain results that are, once again, virtually indistinguishable from ours (see the last two sets of columns in Figure 3).

In other words, Krueger and Zhu

also now report consistently positive results for African-Americans, regardless of how ethnicity is defined, even when kindergartners are included in the analysis—as long as baseline scores (and only baseline scores) are taken into account in the statistical estimation of programmatic effects.

To Ignore or Not to Ignore Baseline Test Scores

Neither changing the definition of African-American nor adding students for whom baseline test scores are missing appreciably changes the results we originally reported. To get different results, still a third methodological step is required. Krueger and Zhu argue that, to avoid a biased estimate, one must ignore baseline test scores, even for those students for whom these are available. But if including baseline scores introduced bias, the magnitude of the effect would change substantially. It does not. Adding baseline scores shifts estimated effects by less than half a percentile point.

Not only is no bias introduced, but including baseline test scores has the advantage of yielding more precise results, allowing researchers to reach firmer conclusions about the efficacy of a programmatic intervention. Estimated impacts from models that control for baseline scores are significant at the .05 level (using the two-tail test), while the less-precise results in models that do not control for baseline scores are significant at only the .10 level, using a one-tail test (a significance level below the threshold Krueger and Zhu find acceptable).

In sum, Krueger and Zhu take three methodological steps to generate results that are not statistically significant: 1) changing the definition of the group to be studied, 2) adding students without baseline test scores, and 3) ignoring the available information on baseline test scores, even though this yields less precise results.

The New York data continue to support the conclusion that disadvantaged African-American students benefit from private schooling.

By contrast, Barnard agreed with our decisions to: 1) use the mother's ethnicity as the basis for defining the child's, 2) focus on those students in the grades for which baseline scores were available for most students, and 3) control for baseline scores, whenever possible. Using pioneering statistical techniques, Barnard reports similar findings, while Krueger and Zhu venture far afield to uncover contrary ones.

The Value of Randomization

Given differences of opinion among researchers, it is easy to jump to the conclusion that randomized field trials are not the gold standard they are thought to be. If social scientists can reach opposite conclusions from the same data set, then research, even from randomized field trials, may do little to inform policy debates.

We take a different view. In New York, the results reported by all parties are consistently positive—only the magnitude of the effects and the level of statistical significance fluctuate. Furthermore, different statistical techniques generate roughly equivalent results. Moreover, the results that we report are consistent with past research on public and private schools. More than 25 years ago, James Coleman and his colleagues found that attending a private school was more beneficial for black students than for whites, as measured by test scores. More recently, Princeton economist Cecilia Rouse, after reviewing the research literature, concluded that "the overall impact of private schools is mixed, [but] it does appear that Catholic schools generate higher test scores for African-Americans." Another literature review, conducted by economists Jeffrey Grogger and Derek Neal, found few clear-cut gains for white students, while "urban minorities in Catholic schools fare much better than similar students in public schools."

Controversies surrounding randomized experiments can nonetheless be reduced by collecting baseline data on the outcome variable of greatest interest—in this case, students' test-score performance. In the absence of this information, experiments devolve into endless arguments over whether random assignment actually occurred and whether the two groups being compared are genuinely equivalent. Consider, for example, the recent skepticism directed toward Tennessee's Project STAR study, a randomized field trial on class size that failed to collect baseline test-score data.

Still, whether or not one restricts the analysis to those cases where baseline test scores were available, results are clear. In New York, private-school attendance positively affected the test scores of African-American students, but not those of any other ethnic group. For this reason, we think that the evidence from New York continues to support the conclusion—also reached in a wide variety of earlier studies—that disadvantaged African-American students living in urban environments benefit from private schooling.

–Paul E. Peterson, the editor-in-chief of Education Next, *and William G. Howell are professors at Harvard University. They are the principal authors of* The Education Gap: Vouchers and Urban Schools *(Brookings, 2002). To view the unabridged version of this article, log on to www.educationnext.org.*